

c) Independence of the data from "external" influences.

As noted above, the data must be as free as possible from the effects of systematic, differential external effects. To include rural installations along with urban installations, for example, will result in biased samples whenever there is a difference in the customer mix served by each provider.

From the above discussion it should be clear that the ability to construct and perform a test which has an exact mathematical confidence level does not in itself provide real confidence in the test result. A thorough understanding of the theoretical requirements of the test and the characteristics of the underlying data is essential to designing such a test.

## **II. USEFULNESS OF TESTS**

### **A. Regional comparisons.**

The use of a standard test might be of some limited use in the measurement of regional differences. However, even with a standard methodology, it would be difficult to determine whether a particular ILEC was providing equal service to CLECs with the same regularity as others. The lack of experience with the data to be gathered will make it very difficult to make these comparisons and the use of standard methods may simply lead to the false expectation that the results should be comparable.

It must be recognized that there will inevitably be limitations in comparing results across ILECs and even across states in the same ILEC's territory. Differences in the method of gathering data and performing service operations will affect the results and may well be specific to a particular company, state or region. Differences in the mix of customers served by CLECs as well as demographic, climatological and topographical differences will always result in external factors that will affect the results in unknown ways.

### **B. ILEC Performance**

Since most ILECs operate in multiple states, the recommendation of standard test methods might reduce the number of reporting formats an ILEC is required to provide. At the same time, state, or even CLEC specific tests, should not be proscribed unless all parties agree that they would be valuable.

Whether or not a standard test methodology is selected, that selection should not limit the possibility of using any other method as part of an investigation into allegations of discrimination. All statistical tests are subject to some probability of error. A good standard test should provide reliable indication of a strong probability of discrimination in the provision of services. These indications should lead to further investigation using all available and appropriate methods, not a summary judgment of guilt.

### **III. WHAT TESTS ARE APPROPRIATE**

#### **A. The Nature of the Problem**

##### **1. Service Comparisons**

There is a predicate matter of defining what “services” are to be measured as between the CLEC and ILEC and among the CLECS themselves. The very nature of competition in the telecommunications industry will lead CLECs to differentiate themselves, in the eyes of a potential customer, from the ILEC and from other CLECs. Each will likely attempt to carve out a niche for itself among all of the potential customers.

Differences in the offerings of these CLECs will cause customers with very different characteristics to select different CLECs or to remain with the ILEC. CLECs with ties to an IXC, for example, might tailor their offerings in a way that attracts heavy users of long distance services. Another, with ties to a cellular provider might target heavy mobile users.

Operationally, these differences will result in CLECs requesting services that do not span the full range of services performed by the ILEC. If a CLEC targets multi-line residential customers, the services it requests will certainly be different from those requested by a competitor who targets single line businesses. The service operations requested by each of them will certainly differ from those performed by the ILEC for customers who continue to purchase its services.

The differences in services requested by these customers will affect the measurements of service quality in as yet unknown ways. Any test used to detect differences in quality levels must recognize the potential for differences resulting from taking measurements across operations that are not equivalent.

## 2. Random Variability

All parties recognize that measurements of service quality, average time to accomplish a task, average time to respond to a call, percent of rejected orders etc., contain random variation. As noted above, the presence of random variation makes it impossible to determine with certainty whether service provided to a CLEC is substantially the same in quality to that provided to other CLECs and to that an ILEC provides itself. For that reason, statistical tests are required.

The best a statistical test can hope to accomplish is to provide evidence as to the probability that there is a difference in quality provided. All statistical tests set a criterion for failure which will result in some percentage of "false alarms," what is usually referred to as a Type I error. That is, the test will indicate the existence of systematically inferior service when there is in fact only random variation. If the number of "false alarms" is high, the usefulness of the tests will be limited. For example, if there were a 5 percent probability of false alarm for each of the 210 separate proposed measurements, there is a 60 percent probability that there will be at least 10 false alarms every month and a 98 percent probability that there will be at least 5 false alarms. A test which continually cries "wolf" has no ability to draw attention to real problems.

One must of course recognize the complementary error, that of not detecting a substantial difference in service when one does exist. Evaluation of the probability of this type of error, usually called a Type II error, first of all requires the specification of the magnitude of a substantial difference. The probability of not detecting an infinitesimal difference certainly is larger than that of not detecting a very large difference. Any accommodation of a test procedure to reduce Type II error therefore must first of all address the question of just how large a difference will have significant negative effects on competition.

## 3. Summary

The design and interpretation of tests to detect substantial differences in the quality of service must recognize the potential for systematic differences resulting from the targeting of customers with different usage and service requirements by different CLECs. In the presence of such self-selection bias, any statistical test which does not control for the source of the bias will be of limited value. The value will be especially low if the results are taken to be a black and white determination of discrimination.

Random variation in service measurements must also be properly considered. If the mathematical confidence level is set too low, resulting in a high probability of false findings of

inequality, the test results will either be ignored or they will simply lead to endless and excessive litigation. The probability of failing to detect real and substantial differences must also be considered, but not at the expense of limiting the test's ability to call attention to areas that require attention.

## **B. Complexity of tests**

The tests used to detect possible differences in service quality should be as simple and understandable as possible. The more complex the test the higher the probability of errors in its execution and the higher the cost of implementation. Simple, well-known tests have their limitations, as I discuss below, but they have the advantage of being at least reasonably well understood. There is no single test that is "best" at detecting all possible differences in service "quality." A well-accepted test that performs reasonably well will be more effective than one or more complex tests which require expert interpretation.

## **C. Test of Central Tendency**

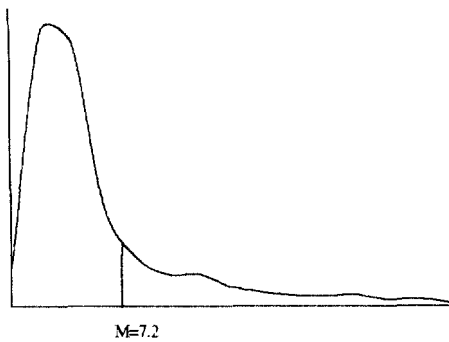
The most important tests that should be performed to assess substantially similar quality of service is one for similarity in the central tendency of samples. Any plausible form of discrimination would almost invariably result in a systematic difference in the central tendency of the measurements. One can, of course, imagine elaborate schemes of discrimination which result in measurements with similar means. The example cited by the FCC is one such example.<sup>3</sup>

There is some disagreement among statisticians with regard to the most appropriate method for testing whether two "samples" were taken from the same "population." In its NPRM, the Commission has stated that the  $t$  test is generally considered the most appropriate test. This is to some extent true. However, one of the key assumptions upon which the  $t$  test is founded is that the samples be drawn from a normal distribution.<sup>4</sup> Seventeen of the twenty-nine measurements proposed in the NPRM are time intervals. All of these intervals are bounded below by zero but may include relatively high values. Their underlying distribution cannot, therefore, be normal. A fundamental assumption of the  $t$  test is therefore violated for more than two thirds of the measurements.

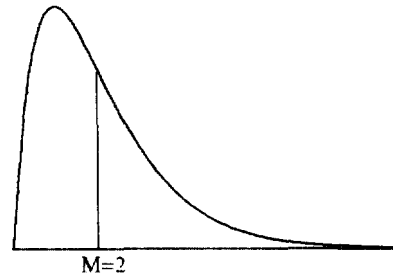
---

<sup>3</sup> FCC-NPRM 98-72. Appendix B, p.2.

<sup>4</sup> Sheskin, David J. "Handbook of Parametric and Nonparametric Statistical Procedures," CRC Press, New York, 1997, p. 153.



**Figure 1: Montana New Business Service Order Interval, days,  
March-December 1997. Source: U S West Communications, Inc.**



**Figure 2: Gamma Distribution.**

As shown in Figure 1, actual data on service completion intervals does not exhibit the familiar symmetric shape characteristic of a normal distribution. The underlying distribution of these measures more closely resembles some form of the gamma distribution (shown in Figure 2), which is often used to model service intervals.<sup>5</sup>

The Z test, for large sample sizes, does not rely on the normality of the underlying distribution. However, its reliability is still affected by the shape of the underlying distribution. The Central Limit Theorem, which forms the theoretical basis for of the Z test, applies unequivocally only to means based on an “infinite” number of samples. Clearly that condition cannot be fully met.

The Z test also assumes knowledge of the variances of the underlying distributions. We are thus left in the situation of having at least two possible tests but not being able to satisfy the theoretical requirements of either one. Fortunately the fact that the measurements that must test do not meet all of the theoretical requirements of the available tests does not mean that these test cannot be usefully applies. It simply requires that we use caution in applying them. For example, if the sample sizes are large, at least some practitioners consider sample variances to be good estimates of the population variances,<sup>6</sup> thus relaxing somewhat the requirement for “known” population variances.

Since the distributional and sample size requirements of the proposed tests are clearly not met, any reasonable statistical test must account for any differences resulting from these

---

<sup>5</sup> Mendenhall, William et al. Mathematical Statistics with Applications, Second Edition. Boston: Duxbury Press, 1981, p. 143.

<sup>6</sup> Mendenhall, ... p. 383.

deviations. Below, we use Monte Carlo simulation methods to show the effect of the deviation of the data from a normal distribution.

#### IV. MONTE CARLO SIMULATION OF TESTS

##### A. Type I Error

A primary consideration in designing a statistical test of substantially similar quality is its Type I error, the probability of a false finding of unequal quality. If our data exactly met all of the criteria for either the  $t$  or the  $Z$  test the Type I error could be calculated exactly. Since that is not the case we must use other methods to make some assessment of the effect on the Type I error of deviations from the required criteria. Both sample size and distribution shape will affect the probability of Type I error. In order to assess the magnitude of that effect I have performed a number of Monte Carlo simulations of the tests under consideration.

The tests were performed by drawing random samples of different sizes from four different distributions. The following four distributions were used in the simulations:

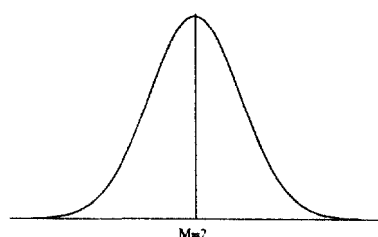


Figure 3: Normal Distribution: Mean of 2 and standard deviation of 1, the common symmetrical "bell" curve

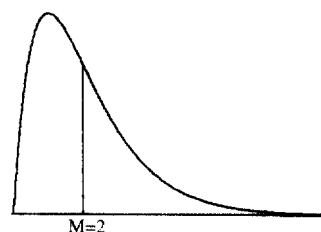
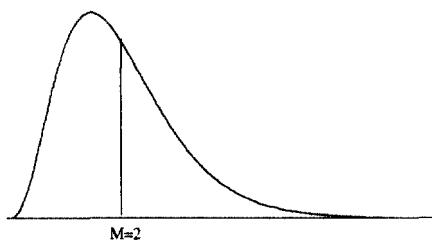
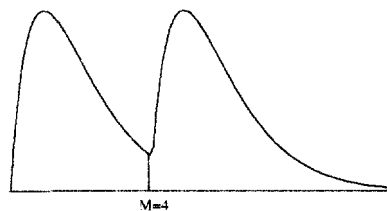


Figure 4: Gamma<sup>1</sup> Distribution: Mean of 2 and standard deviation of 1.4. Bounded on one side by 0.



**Figure 5: Gamma<sup>2</sup> Distribution: Mean of 2 and standard deviation of 2. Bounded on one side by 0.**



**Figure 6: Double Gamma Distribution: Two Gamma<sup>1</sup> Distributions with mean of 4 and standard deviation of 2.5.**

These different distributions represent different possible shapes the underlying data might take. The normal distribution is the common “bell-shaped” curve upon which the *t* and *Z* are based. The distribution labeled Gamma<sup>1</sup>, being bounded on one side by zero, more accurately represents a distribution of time measurements. The distribution labeled Gamma<sup>2</sup> has the same mean as the Gamma<sup>1</sup> distribution but a larger standard deviation. The Double Gamma is constructed from two gamma distributions in order to produce the shape shown. This may approximate situations in which there the underlying population of service operations consists of two distinct distributions, urban and rural installations perhaps.

Samples of fixed size were drawn for an ILEC and samples of varying size for a CLEC. The samples for both were drawn from the same population. Both *Z* and *t* tests were performed at calculated confidence levels of 99 percent, to determine whether a pair of samples failed these tests of equality. The average number of failures that occurred in 100 sets of 1000 sample pairs is reported in the table.

Two methods of calculating the standard deviation of the sampling distribution of the difference in the sample means. The first is the method proposed by the Local Competitor’s Users Group (LCUG) and the second is a pooled estimate.

Equation 1: LCUG Method

$$\sigma_D = \sqrt{\frac{\sigma_{ILEC}^2}{n_{ILEC}} + \frac{\sigma_{CLEC}^2}{n_{CLEC}}}$$

Equation 2: Pooled Method

$$\sigma_D = \sqrt{\frac{\sigma_{ILEC}^2 + \sigma_{CLEC}^2}{n_{ILEC} + n_{CLEC}}}$$

The LCUG method uses only the ILEC sample standard deviation while the pooled method uses both the ILEC and the CLEC sample standard deviation.

A test statistic was calculated, using each of these methods in the following manner to examine whether discrimination has taken place against a CLEC.

$$Z = \frac{\bar{x}_{CLEC} - \bar{x}_{ILEC}}{\sigma_D} \quad \text{Equation 3: Test statistic}$$

The test statistic was compared to a Z-value for 99 percent confidence for the Z-test, and to a *t*-table with 99 percent confidence and  $n_{ILEC} + n_{CLEC} - 2$  degrees of freedom. The results of one set of Monte Carlo simulations are given in Table 1.

1. Normal Distribution.

Since the tests were performed with a 99 percent confidence level, we should theoretically expect approximately 10 test failures in each set of 1000 samples. The mean number of failures found using either method when samples are drawn from a normal distribution is very close to 10 and approaches it even more closely as the CLEC sample size increases.

It is interesting to note that the mean number of failures found by the Z test and the *t* test is very similar for all sample sizes and all distribution shapes. This would seem to indicate that there is little practical difference between the Z and *t* tests at the sample sizes contemplated here.



Tabel 1: Number of Failures for Monte Carlo Simulations  
100 Sets of 1000 Samples,99% Confidence Level

ILEC Sample Size=200									
		CLEC Sample Size=30				CLEC Sample Size=50			
		LCUG		Pooled		LCUG		Pooled	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Normal	Z-test	10.23	3.30	12.41	3.56	10.91	3.26	11.39	2.98
	t-test	9.82	3.20	11.92	3.58	10.51	3.16	10.93	2.98
Gamma <sup>1</sup>	Z-test	18.03	4.29	5.02	2.22	17.73	4.39	6.03	2.32
	t-test	17.37	4.14	4.78	2.14	17.23	4.32	5.76	2.31
Gamma <sup>2</sup>	Z-test	15.90	4.04	6.87	2.69	15.14	3.81	6.98	2.69
	t-test	15.33	3.91	6.59	2.67	14.63	3.75	6.68	2.60
Double Gamma	Z-test	18.05	4.48	4.96	2.23	18.00	3.94	6.53	2.52
	t-test	17.48	4.43	4.69	2.11	17.49	3.81	6.15	2.49
ILEC Sample Size=100									
		CLEC Sample Size=100				CLEC Sample Size=200			
		LCUG		Pooled		LCUG		Pooled	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Normal	Z-test	10.19	2.93	10.20	3.02	10.56	3.01	10.39	2.99
	t-test	9.87	2.79	9.91	2.97	10.27	2.86	10.11	3.00
Gamma <sup>1</sup>	Z-test	16.93	3.71	7.65	2.54	16.25	4.34	9.61	3.08
	t-test	16.32	3.74	7.38	2.44	15.94	4.29	9.37	2.99
Gamma <sup>2</sup>	Z-test	14.46	3.97	8.39	2.96	15.03	3.66	10.33	3.07
	t-test	14.08	4.00	8.06	2.95	14.66	3.67	10.05	3.04
Double Gamma	Z-test	15.89	4.02	7.27	2.61	16.80	4.14	10.40	3.06
	t-test	15.44	3.91	7.07	2.52	16.38	4.00	10.05	2.99

## 2. Gamma Distributions

The results for the samples drawn from all gamma-like distributions diverge significantly from those for the normal distribution. The divergence shows the sensitivity of both the Z and the *t* test to the shape of the underlying distribution. Note also that the direction of the effect is opposite for the two methods of calculating variance. The LCUG method consistently over reports failures whereas the pooled method approaches the theoretic number of failures for larger CLEC sample sizes. The results for the simulations drawn from the Gamma<sup>2</sup> and Double Gamma distributions are very similar showing similar tendencies as CLEC sample sizes increase. As in the case of the normal distribution, the divergence from the theoretical number of failures diminishes as the CLEC sample size increases.

### 3. Effect of sample size on tests

In the preceding table, the ILEC sample size was 200. Since this may not be realistic for each type of test per period, Monte Carlo simulations were conducted for different ILEC sample sizes as well. Table 2 shows the result for ILEC sample sizes of 100 and 50.

These simulations show that for any ILEC and CLEC sample size, if the underlying distribution is normal the number of failures reported by each method is close to the theoretical. The results change dramatically with the gamma distributions. As the ILEC and CLEC sample sizes converge, there is greater accuracy using the pooled estimation of variance, whereas the LCUG method consistently over reports failures. For ILEC and CLEC sample sizes that are different, the pooled method reports less than the theoretically expected number of failures, thus giving a lower Type I error, whereas the LCUG method over reports failures.

[Remainder of page is intentionally blank]

**Table 2: Number of Failures for Monte Carlo Simulations**  
100 Sets of 1000 Samples, 99% Confidence Level

ILEC Sample Size=50									
		CLEC Sample Size=30				CLEC Sample Size=50			
		LCUG		Pooled		LCUG		Pooled	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Normal	Z-test	11.95	3.42	11.91	3.09	12.70	3.37	11.69	3.22
	t-test	10.73	3.17	10.64	3.07	11.68	2.89	10.68	2.99
Gamma <sup>1</sup>	Z-test	25.80	4.61	7.46	2.62	26.13	4.50	10.89	3.13
	t-test	23.94	4.32	6.41	2.61	24.71	4.44	9.79	2.85
Gamma <sup>2</sup>	Z-test	21.01	4.73	8.50	3.16	21.37	5.54	10.81	3.69
	t-test	19.14	4.74	7.49	2.91	19.89	5.28	9.77	3.58
Double Gamma	Z-test	26.23	5.08	7.88	3.01	26.26	4.75	10.47	3.01
	t-test	24.17	4.94	6.75	2.69	24.51	4.64	9.46	2.91

ILEC Sample Size=100									
		CLEC Sample Size=30				CLEC Sample Size=50			
		LCUG		Pooled		LCUG		Pooled	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Normal	Z-test	11.15	2.91	12.07	3.36	10.90	3.07	10.93	2.99
	t-test	10.27	2.71	11.14	3.34	10.17	2.94	10.34	2.90
Gamma <sup>1</sup>	Z-test	19.86	4.89	5.79	2.56	20.74	4.83	7.32	2.39
	t-test	18.66	4.51	5.19	2.36	19.89	4.76	6.75	2.38
Gamma <sup>2</sup>	Z-test	17.26	4.38	6.70	2.56	17.33	3.79	8.60	2.70
	t-test	16.16	4.15	6.29	2.51	16.48	3.65	8.15	2.57
Double Gamma	Z-test	20.55	4.23	5.57	2.08	19.89	4.33	7.39	2.63
	t-test	19.37	3.96	5.09	1.93	18.96	4.25	6.91	2.52

ILEC Sample Size=200									
		CLEC Sample Size=100				CLEC Sample Size=200			
		LCUG		Pooled		LCUG		Pooled	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Normal	Z-test	11.10	3.23	10.44	3.28	10.86	3.39	10.55	3.39
	t-test	10.60	3.15	9.93	3.26	10.45	3.25	10.12	3.40
Gamma <sup>1</sup>	Z-test	19.92	4.55	10.18	3.52	19.91	4.92	13.27	4.12
	t-test	19.24	4.51	9.63	3.51	19.46	4.96	12.73	3.95
Gamma <sup>2</sup>	Z-test	17.98	3.98	10.62	3.26	16.78	3.97	12.49	3.16
	t-test	17.27	3.89	9.97	3.22	16.41	3.96	12.09	3.21
Double Gamma	Z-test	20.39	4.60	10.38	3.08	20.05	4.21	13.49	3.45
	t-test	19.50	4.50	9.87	2.92	19.66	4.16	13.07	3.37

The results of the Monte Carlo simulations show consistently that the LCUG variance calculation method gives higher than expected numbers of failures for gamma-like distributions. Even when ILEC and CLEC sample sizes are the same, the number of failures generated though the LCUG method of calculating variance is higher than the pooled method of calculating

variance. This affect arises because a failure is the result of two factors: a relatively high difference between CLEC and ILEC means and relatively low variances (see Equation 3). Since the difference in means is calculated as the CLEC mean minus the ILEC mean, a large difference implies a high CLEC mean. For gamma distributions, samples with high means usually have a higher variance, and those with low means have a lower variance. Thus, when a difference is great enough to warrant a failure, the CLEC variance is usually high and the ILEC variance low. The low ILEC variance is used twice in the calculation of Z for the LCUG method, and therefore increases the Z-value such that there are a greater number of failures. Conversely, in the pooled method, the high CLEC variance balances the low ILEC variance, making the number of failures lower.

The results of the test show, therefore, that when sample sizes are equal, the pooled method of calculating variance is more accurate in determining failures for gamma distributions. However, it must be cautioned that each of these methods of calculating variance has its shortcomings, depending on sample sizes and the underlying distribution of the sample. Thus, it is important to recognize that no one single test may be appropriate in all cases of potential discrimination, and that each situation must be thoroughly investigated before discrimination may be charged.

#### 4. Summary

The Monte Carlo simulations show very clearly that either the Z or the *t* test will produce similar results with combined test sizes greater than 100. With smaller sample sizes some adjustment would be required in the confidence level in order to realize the desired confidence level. Failures reported on samples from normal distributions are consistent with theory regardless of the testing method. However with gamma-like distributions, tests using the LCUG method consistently over report failures for the aforementioned reasons. Tests using the pooled method report fewer failures when ILEC and CLEC sample sizes are different. As the sample sizes converge, the pooled method more accurately reports failures.

Although these simulations provide valuable information, they should not be considered definitive. The results should be interpreted as an indication of the degree of divergence from theory one can expect with different distributions and combinations of sample size. They also lend much support to the use of a minimum CLEC sample size of about 50, a compromise between reality and the 30 referred to by the FCC in its NPRM, and commonly proposed by statistical texts. Based on these results it is clear that relying on calculated confidence levels in designing a

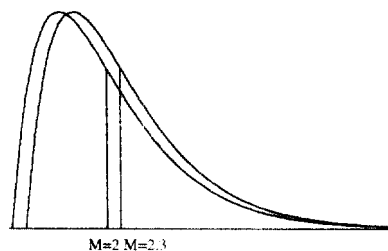
test for discriminatory service can lead to serious errors. In order to provide consistent probabilities of Type I error, tests must be designed for the specific data at hand.

## **B. Ability to detect inequalities in service**

The above simulations show that the sample size, the underlying distribution, and the method used to calculate the variance affects the realized Type I error of a test. Type II error is the probability of not finding inequality of service when it exists. Given that fact, it is reasonable to expect that the ability to detect specific types of inequality will also be affected by the same factors.

In order to gain some insight into the magnitude of this effect on different distributions, another set of Monte Carlo simulations was performed. Three different scenarios of discrimination were simulated in order to examine the effect of the variance methods and sample size on detection. We assumed that the CLEC samples would be taken from the following distributions whereas the ILEC sample would be taken from the Gamma<sup>1</sup> distribution described above.

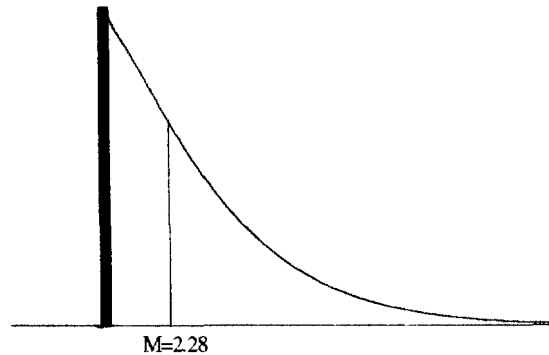
### **1. Difference in mean, equal variance**



**Figure 7: Staggered Discrimination. CLEC mean of 2.3, standard deviation of 1.4.**

In this case the ILEC and CLEC distributions are as shown in Figure 7. These distributions differ only in their mean, not in their variance. This would be the result of a discriminatory policy that added the same interval of time to every CLEC order.

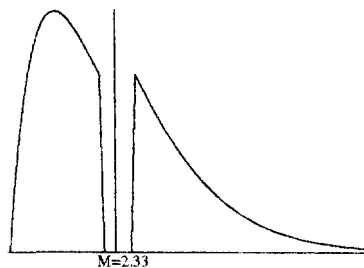
2. **Difference in mean, CLEC variance smaller than ILEC variance**



**Figure 8: Truncated Distribution. Mean of 2.28  
Standard deviation 1.2**

Figure 8 shows the case where the CLEC variance is smaller than the ILEC variance, when for example an ILEC may be setting a minimum time to complete CLEC orders. To produce this distribution, any number taken from the Gamma<sup>1</sup> distribution that was lower than 1.5 was converted to 1.5. In this case, the mean of the CLEC distribution has been increased by the same degree as in the former case, but the variance is smaller.

3. **Difference in mean, CLEC variance larger than ILEC variance**



**Figure 9: Split Distribution. Mean of 2.33,  
standard deviation of 1.7**

This distribution might result from an ILEC following a policy which causes only some CLEC orders to require longer intervals, with others being performed in a nondiscriminatory manner. Here, any orders that are require more than 2 days are delayed by .8 days. This

effectively increases the variance of the distribution, but the difference in mean is similar to the above two scenarios.

The results of Monte Carlo simulations are compared to expected values calculated from the given confidence intervals in the following table:

**Table 3: Number of Failures for Monte Carlo Simulations**  
100 Sets of 1000 Samples, 99% Confidence Level  
ILEC Sample Size=200

	CLEC Sample Size=30					CLEC Sample Size=50				
	Expected	LCUG		Pooled		Expected	LCUG		Pooled	
		Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.
Stagger	108	123	10.69	78	8.46	160	181	11.93	139	10.94
Truncated	56	82	7.48	76	7.69	101	132	12.00	152	12.45
Split	174	182	13.27	67	8.35	239	238	11.71	112	9.40
	CLEC Sample Size=100					CLEC Sample Size=200				
	Expected	LCUG		Pooled		Expected	LCUG		Pooled	
		Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.
Stagger	278	292	15.15	268	15.88	421	427	16.53	420	16.00
Truncated	208	239	13.14	295	13.73	253	374	15.08	437	14.99
Split	360	356	15.27	227	13.94	505	494	15.22	392	15.98

In all cases, larger sample size is the most important factor in the detection of discrimination. The simulations show that a pooled variance is more effective in the Truncated distribution, which produces smaller variances. For the Split distribution, the LCUG method of determining variance detects discrimination better than the pooled method.

#### **E. Conclusion**

This series of tests shows that the pooled variance renders the test more sensitive to situations in which the CLEC variance is smaller than the ILEC variance and less sensitive to situations in which the CLEC variance is higher than the ILEC variance. Since most scenarios of covert discrimination against a CLEC will reduce variance of the CLEC sample, the pooled variance seems most appropriate.

## V. AT&T'S EX PARTE SUBMISSION

The *ex parte* submission of AT&T contains many misleading statements.<sup>7</sup> The following is an analysis of their proposed three criteria for failure.

### 1. Maximum number of comparisons failing the test.

In their example they state that if one uses a 95 confidence level there should be no more than 5 comparisons in 100 that fail the test. In fact, given a five percent probability of failure, the probability of finding at least six failures in 100 tests is 38 percent. The use of this criteria would therefore result in a false finding of discrimination more than one third of the time if there were 100 comparisons made each month. The fallacy here is that one cannot translate long run probabilities into probabilities for finite samples. If I toss a fair coin twice and get two heads, there is no law of nature which says that the next two tosses must be tails in order to "catch up" with the long run probability which indicates that I should have two heads and two tails.<sup>8</sup>

### 2. Maximum repeating measurements failing the test

This criteria is also very misleading. Even though the probability of a single comparison failing the test in two months is 0.0025, the probability of finding at least one such test in 100 is 22 percent. This sounds unbelievable until one takes a close look at the probabilities. First of all the probability that a particular test will not have a repeat failure is  $1 - 0.0025$  or 0.9975. In order for there to be no repeat failures in 100 tests there must be 100 tests with zero repeat failures. The probability of this event is the probability that one test will not repeat raised to the 100 power. That is  $0.9975^{100}$  or 0.7786 and its complement, the probability that there will be more than zero with repeat failures is  $1 - 0.7786$  or 0.2214.

### 3. Measurements exhibiting extreme differences

The criteria described here assumes that the CLEC is entitled to superior service. My understanding is that such is not the appropriate legal standard. Furthermore, as noted above, service quality measurements contain a substantial random component and the probability of an

---

<sup>7</sup> Referenced in FCC-NPRM 98-72 Appendix B, p.3.

<sup>8</sup> For an interesting discussion of this phenomenon see Stewart, Ian "Repealing the Law of Averages." Scientific American, April 1998, 102.



event having a measurement greater than three standard deviations, although very unlikely, is not impossible.


## **VI. CONCLUSION**

Although the thrust of my comments has been to pointing out many of the shortcomings of the statistical tests under consideration, I continue to support the use of such tests. The messages that should be taken from the above demonstrations are as follows:

- A. Statistical tests can be useful indicators of substantial differences in the quality of service operations.
- B. Any test must recognize and accommodate the characteristics of the actual data, not some convenient theoretical distribution.
- C. Test results should be used as indicators only not presumptive measures of discrimination. Careful examination of the data and more sophisticated tests are required to make any such determination.

## CERTIFICATE OF SERVICE

I, Kelseau Powe, Jr., do hereby certify that on this 2<sup>nd</sup> day of June, 1998, I have caused a copy of the foregoing **MOTION TO ACCEPT LATE-FILED COMMENTS OF U S WEST COMMUNICATIONS, INC. and COMMENTS OF U S WEST COMMUNICATIONS, INC.** to be served, via hand delivery, upon the persons listed on the attached service list.

  
Kelseau Powe, Jr.

William E. Kennard  
Federal Communications Commission  
Room 814  
1919 M Street, N.W.  
Washington, DC 20554

Gloria Tristani  
Federal Communications Commission  
Room 826  
1919 M Street, N.W.  
Washington, DC 20554

Michael K. Powell  
Federal Communications Commission  
Room 844  
1919 M Street, N.W.  
Washington, DC 20554

Harold Furchtgott-Roth  
Federal Communications Commission  
Room 802  
1919 M Street, N.W.  
Washington, DC 20554

Susan P. Ness  
Federal Communications Commission  
Room 832  
1919 M Street, N.W.  
Washington, DC 20554

Chief, Common Carrier Bureau  
Federal Communications Commission  
Room 500  
1919 M Street, N.W.  
Washington, DC 20554

Carol E. Matthey  
Federal Communications Commission  
Room 544  
1919 M Street, N.W.  
Washington, DC 20554

Janice M. Myles  
Federal Communications Commission  
Room 544  
1919 M Street, N.W.  
Washington, DC 20554

**(Including 3x5 inch diskette, w/cover letter)**

Brent Olson  
Federal Communications Commission  
Room 544-K  
1919 M Street, N.W.  
Washington, DC 20554

Radhika Karmarkar  
Federal Communications Commission  
Room 544  
1919 M Street, N.W.  
Washington, DC 20554

International Transcription  
Services, Inc.  
1231 20<sup>th</sup> Street, N.W.  
Washington, DC 20036

(CC98-56a.doc)  
Last Update: 6/2/98